AMENDMENTS TO THE SPECIFICATION

Please amend the specification as follows:

Page 9, please amend the paragraphs at lines 7-8 to read as follows:

Fig. 3 is a reference histogram of the data accesses of Figs. 2A-2C; ~~and~~

Fig. 4 is a diagram showing the reference affinities of various fields in tree form; and

Figs. 5A-5F are flow charts showing various processes performed in the preferred embodiment.

Page 13, between lines 5 and 6, please insert the following new paragraph:

The process described above will be summarized with reference to the flow chart of Fig. 5A. In step 5001, a last access time of each of the data is determined. In step 5003, a search tree is organized from the last accesses, wherein the search tree comprises a node for each of the data, the node comprising the last access time and a weight of a sub-tree of the node. In step 5003, the search tree is compressed in accordance with the bounded relative error.

Page 13, between lines 20 and 21, please insert the following new paragraph:

The process described above will be summarized with reference to the flow chart of Fig. 5B. In step 5101, the last access time of each of the data is determined. In step 5103, a trace is maintained storing the last access times of the last C accesses of the data. In step 5105, a search tree is maintained storing access times other than the last C accesses, each node in the search tree having a capacity B.

Page 15, please amend the paragraph at lines 11-16 to read as follows:

As shown in the flow chart of Fig. 5C, a reuse distance histogram is formed in step 5201. Given two reference histograms from two different data inputs (called training inputs; see step 5203), one can construct a formula for each bin. Let d_{1i} be the distance of the ith bin in the first histogram, d_{2i} be the distance of the ith bin in the second histogram, s_1 be the data size of the first training input, and s_2 the data size of the second input. Linear fitting is used in step 5205 to find the closest linear function that maps data size to reuse distance. Specifically, the task is to find the two coefficients, c_i and e_i, that satisfy the following two equations.

Page 26, please amend the paragraph at lines 11-14 to read as follows:

A related technique is called *k*-percent analysis, or *k* % clustering. That technique, as shown in Fig. 5F, compares the reuse signatures of the data in step 5501 and groups two reuse signatures *X* and *Y* (of length *B*) in step 5503 if the difference *p,* given by the equation below, is less than *k*%. The difference in each bin, $|x_i - y_i|$, can be the number of reuses, the sum of the reuse distance, or both.

Page 27, between lines 17 and 18, please insert the following new paragraph:

The above will be summarized with reference to Fig. 5D. In step 5301, a reuse distance histogram is formed. Ins step 5303, from the reuse distance histogram, an affinity group is formed of at least two data which are always accessed within a distance *k* of one another, wherein *k* is a predetermined quantity. In step 5305, the data are selected in the affinity group such that the data in the affinity group have average reuse distances which fulfill a necessary condition with respect to *k*.

Page 27, please amend the paragraph at lines 18-23 to read as follows:

In practice, a stricter condition is used to build a group incrementally, as will be explained with reference to Fig. 5E. Initially, in step 5401, each data set is a group. Then the groups are traversed in step 5403, and two groups are merged if a member in one group and another member in the other group satisfy Equation 1. The process terminates in step 5407 if it is determined in step 5405 that no more groups can be merged. The distance difference is calculated between any two data sets in $O(g^2)$ time. The iterative solutions takes at most $O(g^2)$. The incremental solution takes linear time if implemented using a work-list.